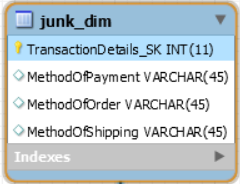
Timeline

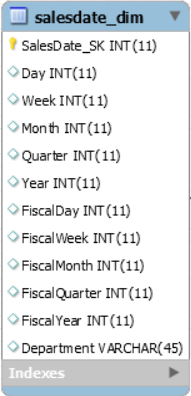
Description automatically generated

The above data mart of all the divisions consist of 6 dimensions viz. Product\_dim, Customer\_dim, Supplier\_dim, orderdate\_dim, salesdate\_dim and 1 fact table sales\_fact table

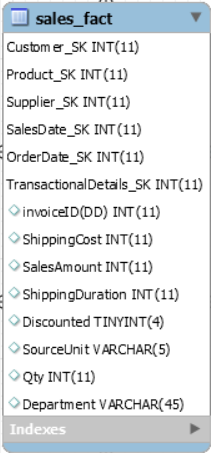
The fact table consist of measurable facts like ShippingCost, ShippingDuration, Discounted, SourceUnit, Quantity, Sales\_Amount

****

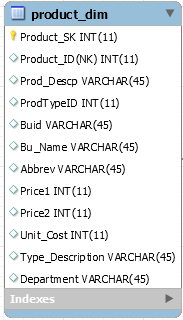
**Junk\_Dimension**

****

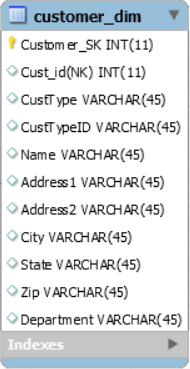
**Sales\_date Dimension**



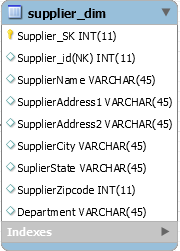
**Sales\_Fact Dimension table**



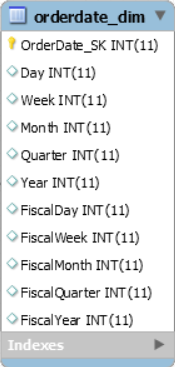
**Product\_Dimesion**

****

**Customer\_Dimesion table**



**Supplier Dimension table**



**Order\_Date Dimension**

|  |  |  |  |
| --- | --- | --- | --- |
| DM Table | Attribute | Problem | Resolution Strategy  (attach code) |
| PECcustomer | Custtype | The data was enclosed in double-quotes  The data was not uniform  Ex: SATELOCALGOVT, USGOVT,COMERCIAL | The double quotes where replaced with space  Replaced the ununiform data to maintain data uniformity throughout the data mart |
| PECcustomer | Address | The data was too big in one attribute  The data was not uniform  Ex: St, Street;  Rd. Road; Ave. Avenue | Added Comma delimiter in address field to split into addr1 & addr2 in csv file of customer.  Replaced the data where necessary to have consistent data  Made Rd. to Road  St. to Street  Ave. to Avenue |
| PECcustomer | Name | Partial records had Inc. while some had Incorporation  Similarly some data had Co. while some had Company | Standardized the name,  Inc to Inc.  Co. to Company  Corp to Corporation by replace in string method in pentaho. |
| PECcustomer | Add1 & Addr2 | Some records had St, Ave and Rd etc. so standardized them | Standardized the addr1 & addr2,  Rd. to Road  St. to Street  Dr. to Drive  Av. To Avenue  Ave. to Avenue  by using replace in string in pentaho. |
| TPCWcustomer | All attributes | All attributes- values where enclosed in double quotes | Replaced double quotes & Comma with space in the fields using replace in string in pentaho. |
| TPCWCcustomer | State | The records where inconsistent with PEC state | Standardized all the state names with their ABBREV using replace in string in pentaho. |
| TPCWcustomer | CustType | Records where inconsistent some had Gov ,Comm. Edu so | Standardized the data from Govt to US\_Govt  State to State\_Local\_Govt  Comm to Commercial  Edu to Education  using replace in string in pentaho. |
| TPCWcustomer | Address | The address field was not consistent | Added Comma delimiter in address field to split into addr1 & addr2 |
| TPCWcustomer | Name | Records where inconsistent  Some had Inc, Corp\*, Cor, Chem | Standardized the name, Inc. to Inc, Co. to Company, Corp to Corporation, Comm to Commercial, Chem to chemicals  Firstfed America BanCompany to Firstfed America Ban Company, YuliChem to Yuli  Chemicals  using replace in string in pentaho. |
| TPCWcustomer | Address | Some records had St, Ave and Rd,Dr., etc. so standardized them | Standardized the addr1 & addr2,  Rd. to Road  St. to Street  Dr. to Drive  Av. to Avenue  Ave. to Avenue  & Replaced all . to space  using replace in string in pentaho. |
| TPEcustomer | Name | Records had inconsistency Co., Corp, Inc,Vew etc | Standardized the data from  Inc. to Inc  Co. to Company  Corp to Corporation  Vec to V.e.v  And removed “ .” and replaced with space by using replace in string in pentaho. |
| PECbusinessunit.csv | ABBREV | The data was quoted in doubled quotes  And Had null values | Added missing value  And replaced double quotes(“) with space |
| TPCWbusiness\_unit.csv | ABBREV | Null value | Added missing value |
| TPCEbusinessunit.csv | ABBREV | Null value  And Data inconsistency like  Supply, chemical | Added missing value  & replace Supply with Supplies  & replaced Chemical with Chemicals |
| PECinvoice | All Fields except orderDate | Remove | Select orderDate by using replace in string in pentaho to separate orderDate from the csv file. |
| PECInvoice | OrderDate | Standardization | Standardization the date by replacing ‘-‘ &’.’ by ‘/’. |
| PECInvoice | OrderDate | Standardization | Standardizaring the date format to mm/dd/yy |
| PECInvoice | OrderDate | Extract date, month, year | Extract date, week, quarter month, year from given date using calculator function of pentaho |
| PECInvoice | OrderDate | Calculating Fiscal day, fiscal week, fiscal month, fiscal year | Writing a javaScript to calculate fiscal day, fiscal month, fiscal year, fiscal week, fiscal quarter |
| PECinvoice | All Fields except saleDate | Remove | Select SalesDate by using replace in string in pentaho to separate SalesaDate from the csv file. |
| PECinvoice.csv | saleDate | Standardization | Standardization the date by removing ‘-‘, ‘.’ and adding ‘/’. |
| PECinvoice.csv | All fields | Standardization | Standardization records which includes some random data. |
| TPCWinvoice.csv | saleDate | Standardization | Standardization the date by removing ‘-‘ , ‘.’, ‘,’ by ‘/’. |
| TPCWinvoice.csv | saleDate | Standardization | Standardization the date to mm/dd/yyyy |
| PECinvoice  TPCWinvoice  TPCEinvoice | CalendarYear  CalendarQuarter  CalendarMonth  CalendarWeek  CalendarDay | Extract Year of date A  Quarter of date A  Month of date A  Week of date A  Day of month of date A | Splitting SalesDate and using calculator in pentaho into  CalendarYear  CalendarQuarter  CalendarMonth  CalendarWeek  CalendarDay |
| PECinvoice  TPCWinvoice  TPCEinvoice | CalendarYear  CalendarQuarter  CalendarMonth  CalendarWeek  CalendarDay | Calculate fiscal Year, fiscal week, fiscal month, fiscal quarter | Calculating  fiscal Year, fiscal week, fiscal month, fiscal quarter  from date using Modified JavaScript |
| TPCWinvoice.csv | Custid | Standardization | Negative values were changed to absolute values. |
| SupplierTPCE | NAME  ADDR1  ADDR2  CITY  STATE  ZIP  SUPPLIERID | Standardization | Standardized these attributes using select values to  supplierName  supplierAddress1  supplierAddress2  supplierState  supplierCity  suppplierZipcode  supplier\_Id |
| ProductTPCW | Field\_000  Field\_001  Field\_002  Field\_003  Field\_004  Field\_005  Field\_006  Field\_007  Field\_008  Field\_009  Field\_0010 | Standarization | Extracted all the values and renamed them to  Prodid  ProdDescription  Price1  Price2  UnbitCXost  SupplierNBame  Addr1  Addr2  City  Zip  ProdType\_Id |
| ProductTPCW | supplierCityState | Combined | Splitted city and state using split fields twith delimiter as’,’  supplierCity  supplierState |
| ProductTPCW | ProdId  State  ProdTypeID  City | Enclcosed in quotes | Replacing quotes with space |
| ProductTPCW | supplierName  supplierAddr1  supplierAddr2  supplierState  supplierCity  suppplierZipcode  supplierID | Duplicate Rows | Duplicate rows were removed using Unique rows (Hash Set) |
| ProductTPCW | supplierAddr1 | Extra attribute Attn : | Attn: was removed] before all address 1 in supplier were using replace in string. |
| ProductPEC | prodid  prodDescription  productTypeID | Quotes | Quotes where replaced and some values where changed like Equip\* to Equipment using Replace in string |
| TPC\_W\_Product.csv | Product name | Some records had Equip | Hence replaced with Equipment using replace string (Using Regular expression). |
| TPCWProduct.csv | Unit cost | Values were null | Calculated unit Cost for PEC data using the formula: Total Manufacturing Cost/ Total Quantity, from the manufacturing\_cost.csv and pec.invoice.csv file. |
| TPC\_W\_Product\_Type.csv | All fields | The data contains double quotes””. | Removed quotes using Select Value in pentaho |
| TPC\_W\_ Business\_ Unit.csv | ABBREV | Standardizing | Added missing data and formatted exiting data for uniformity like Supply to supplies. |
| TPC\_E\_Product.csv | Description | Some of the description contains Equip. | Hence replaced with Equipment using replace string |
| TPC\_E\_Product\_Type.csv | Type description | Some of the type description contains Equip. | Hence replaced with Equipment using replace string (Using Regular expression). |
| TPC\_E\_Product\_Type.csv | BUID | Data column contains double quotes ““ | Hence removed using string operations. |
| TPC\_E\_Business Unit.csv | ABBREV | Supply and Chemical is present. | So, replaced it with Supplies and Chemicals using replace string |
| PEC\_Product.csv | Missing Source column | Missing source column | Created a new source column to recognize the source of every record from which company it has been collected. |
| PEC\_Product\_Type.csv | All fields | The data contains double quotes””. | Hence removed using replace in string. |
| PEC\_Product\_Type.csv | Type description | Some of the type description contains Equip. | Hence replaced with Equipment using replace string (Using Regular expression). |
| PEC\_ Business\_ Unit.csv | ABBREV | Some of the miscellaneous data is missing. | Hence added using replace string. |
| Sales\_Fact\_Table  (PEC\_Product) | All Fields | The Values were enclosed in double quotes. | Replaced double quotes with space in the fields using replace in string in pentaho. |
| Sales\_Fact\_Table  (PEC\_Invoice) | Missing Source column | Missing source column | Created a new source column to recognize the source of every record from which company it has been collected. |
| Sales\_Fact\_Table  (TPC\_W\_Product) | Field\_000  Field\_001  Field\_002  Field\_003  Field\_004  Field\_005  Field\_006  Field\_007  Field\_008  Field\_009 |  | The header row names were missing for every column hence replaced the name with the appropriate headed name and renamed them to  Product ID  Product Name  Price 1  Price 2  Unit Cost  ProdTypeID  supplierName  supplierAddr1  supplierAddr2  supplierCityState  supplierZipcode |
| Sales\_Fact\_Table  (TPC\_W\_Invoice) | All Fields | The Values were enclosed in double quotes. | Replaced double quotes with space in the fields using replace in string in pentaho. |
| Sales\_Fact\_Table  (TPC\_E\_Invoice) | Missing Source column | Missing source column | Created a new source column to recognize the source of every record from which company it has been collected. |

|  |  |  |
| --- | --- | --- |
| DM table | Image Creation process | Attach code |
| Product\_Dimension | 1. Extracted the Product, Product\_type and Business\_unit for all the sources and manufacturing cost for PEC 2. Calculated unit cost for PEC data using the formula: Total Manufacturing Cost/ Total Quantity, from the manufacturing\_cost.csv and pec.invoice.csv file. 3. Appended the calculated unitCost in PEC where the unitCost value is null 4. Add Source (for TPCE TPCW and PEC) 5. Merged product, product\_type and BUID for all sources. 6. Add null Record 7. Added surrogate key and send it to output file. | **Product\_Dimension.ktr**  ***Refer to the images below for transformation*** |
| Transaction\_Profile\_  Dimension | 1. Extracted PEC invoice csv file 2. Added DivisionID (1 for TPCE, 2 for TPCW and 3 for PEC) 3. Merge all division Data 4. As TPCW and TPCE does not have any values, added 2 rows with all attributes having initial value 0. One row for TPCW and another row for TPCE. 5. Add null Record 6. Added surrogate key and sent it to output file. | **Transaction\_Profile\_**  **Dimension.ktr**  ***Refer to the images below for transformation*** |
| Supplier\_Dimension | 1. Extracted Supplier from TPE, Product TPCW and productPEC. 2. Added source to all the files to get track of data from divisions. 3. Standardized the format of data from supplier TPC E.   NAME to supplierName  ADDR1 to supplierAddr1  ADDR2 to supplierAddr2  CITY to city  STATE to state  Zip to supplierZip   1. Removed ProductID , ProductName, Price1, Price2, UnitCost,PRODTYPEID 2. Splitted citystate to city and state 3. Updated all mis-spelled address and city and state. 4. Sorted the data from TPC W and TPC E to according to supplierName and merge them. 5. After merging all data removed extra column supplierName\_1, supplierAddr1, supplierAddr2,   City\_1, state\_1, supplierZipcode   1. Removed prodid, prodDescription, 2. Price1, price2, UnitCost, productTypeID from productPEC. 3. Removed all null values using filter rows. 4. Added surrogate key using add sequence. 5. Added supplierAddress1, supplierAddress2, city, state and zip using add constants. 6. Appended all the values from TPCE and PEC. 7. Generated supplier name, supplier address, city, state anfd zip and appended those values with previous append. 8. Added surrogate key. 9. Added SCD using add constant. 10. Taking table output using table output. | **Supplier\_Dimension.ktr**  ***Refer to the images below for transformation*** |
| Customer\_Dimension | 1. Extract customer and customer\_type files for all three divisions PEC, TPCW,TPCE 2. Add ADDR1 & ADDR2 column to TPCW and PEC 3. Map states to its abbreviation in TPCW 4. Replace the abbreviations such as Ave, Rd, Dr in ADDR 2 with complete name in all divisions 5. Add DivisionID (1 for TPCE, 2 for TPCW and 3 for PEC) 6. Merge Customer and Customer\_type of each division 7. Merge all the three division tables 8. Add Null value 9. Add surrogate keys and send it to table output file. | **Customer\_Dimension.ktr**  ***Refer to the images below for transformation*** |
| Order\_Date\_Dimension | 1. Extract PECinvoice.csv file 2. Convert orderDate into MM/dd/yyyy format. 3. Removed all other attributes and only kept orderDate 4. Used Unique hashset to take unique records of saleDate 5. Add fiscal year, fiscal month, fiscal week, and fiscal quarter to the orderDate. 6. Add null records 7. Add surrogate keys and send it to table output file. | **Order\_Date\_Dimension.ktr**  ***Refer to the images below for transformation*** |
| Sales\_Date\_Dimension | 1. Convert the sale date into MM/dd/yyyy format. 2. Removed all other attributes and only kept orderDate 3. Add fiscal year, fiscal month, fiscal week and fiscal quarter to the salesDate. 4. Merge all the 3 divisions 5. Used Unique hashset to take unique records of saleDate 6. Add null Record 7. Add surrogate keys and send it to table output file. | **Sale\_Date\_Dimension**  **.ktr**  ***Refer to the images below for transformation*** |
| Sales\_Fact\_Dimension | 1. Extracted the Product, invoice files from the three sources that is PEC, TPC\_W and TPC\_E 2. Changes the price data type to number. 3. Sorted all the rows with respect to prodID for the sources. 4. Then merged the TPC\_W\_invoice and TPC\_W product on prodID 5. Then merged the TPC\_E\_invoice and TPC\_E product on prodID 6. Then merged the TPC\_E\_invoice and TPC\_E product on prodID 7. Then calculated the number of days to ship by subtracting order date from sales date using the calculator only for PEC data this is because the order date and sales date are present only in this source. 8. Total sales amount was calculated using java script using the formula.   If discount = 0 then we multiply price 1 and quantity, if discount is 1 then we multiply price 2 and quantity.   1. Then the invoice was remned to invoiceID. 2. Post this unwanted attribute for this fact table were removed using select value function. 3. Then all the data was appended for the three sources. 4. Finally given to the table output. | **Transcation\_Profile\_Dimension.ktr**  ***Refer to the images below for transformation*** |

|  |  |
| --- | --- |
| **DM Table** | **Image Creation Process** (attach code) |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |